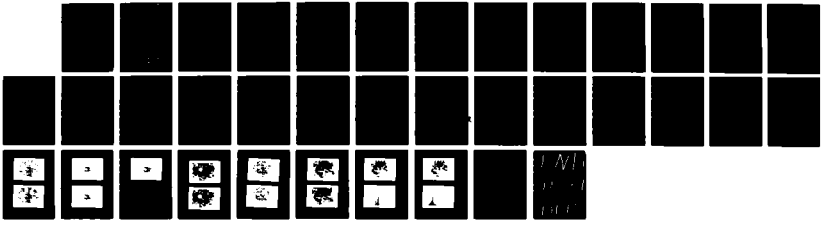
AD-A105 002 STRUCTURE FROM MOTION(U) MINNESOTA UNIV DULUTH
W B THOMPSON NOV 76 AFOSR-TR-87-1576 $AFOSR-85-0382

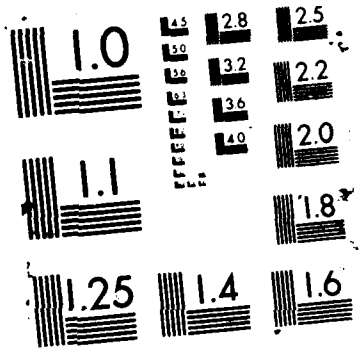
1/1

UNCLASSIFIED

F/G 12/9

NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION	1b RESTRICTIVE MARKINGS
AD-A185 802 (S)	3 DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
	5 MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 87-1576
6a NAME OF PERFORMING ORGANIZATION: University of Minnesota-Duluth	6b OFFICE SYMBOL (If applicable):
6c ADDRESS (City, State, and ZIP Code): Duluth, MN 55812	7a NAME OF MONITORING ORGANIZATION: AFOSR/NM
	7b ADDRESS (City, State, and ZIP Code): AFOSR/NM Bolling AFB DC 20332-6448
8a NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR	8b OFFICE SYMBOL (If applicable): NM
	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: AFOSR-85-0382
8c ADDRESS (City, State, and ZIP Code): AFOSR/NM Bolling AFB DC 20332-6448	10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. 61102F; PROJECT NO 2304; TASK NO A3; WORK UNIT ACCESSION NO

11 TITLE (Include Security Classification): Structure From Motion

12 PERSONAL AUTHOR(S): Professor Thompson

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Final	FROM 9/30/85 TO 11/30/86	11/86	22

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Significant results were obtained on the problems associated with motion-based segmentation. A method for combining motion-based edge detection techniques has been devised. Also, the interpretation of the structure of motion boundaries has been investigated in human vision. Papers have included such titles as "Relative motion: Kinetic information for the order of depth at an edge", and "Acceleration-based structure from motion".

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	
22a NAME OF RESPONSIBLE INDIVIDUAL: Maj. John Thomas	22b TELEPHONE (Include Area Code): (202) 767-5026 / 22c OFFICE SYMBOL: NM

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED

FINAL REPORT - STRUCTURE FROM MOTION
AFOSR Contract AFOSR-85-0382

a. Objectives.

Our principal objective continues to be the development of a robust computational approach for estimating the spatial organization of a scene using time varying properties of image sequences. Under this contract, we have been investigating improved methods for interpreting optical flow from image sequences. Emphasis is placed both on what spatial properties should be computed and on appropriate computational architectures for accomplishing this task.

Two research questions have been investigated over the last year.

Interpreting optical flow at object boundaries.

How can the analysis of optical flow be used to detect object boundaries? How can the three-dimensional structure of object boundaries be determined based on optical flow? How can motion and non-motion information be integrated into more reliable detection and interpretation processes? The principal objective here is to work towards the development of *motion-based segmentation* techniques for image understanding. Motion-based segmentation has the potential not only for locating object boundaries, but also for reducing problems due to occlusion and for providing three-dimensional information useful for object identification and analysis.

Robust methods for determining object motion.

How can the motion of object relative to the camera be determined in a robust manner? The principal objective over the last year has been to develop techniques for detecting moving objects. This is a difficult task when the camera is also moving and the goal is to detect objects moving with respect to the environment, not the camera. Work has also proceeded on new methods for estimating the parameters of object motion.

b. Status of research effort.

Interpreting optical flow at object boundaries.

Significant results have been achieved on the problems associated with *motion-based segmentation*. Discontinuities in optical flow are necessarily due to surface boundaries or discontinuities in depth in the scene. Thus, detected edges in flow necessarily correspond to important properties of scene geometry, where as edges in properties such as luminance can be due to a wide variety of scene properties. Our approach is based on understanding the three-dimensional scene structure leading to an edge in optical flow. As a result, we can simultaneously detect edges and determine important three-

dimensional properties of the associated scene surfaces.

Two significant accomplishments have been achieved on this problem during the last year. We have developed a method for combining motion-based edge analysis with more traditional edge detection techniques. This integrated approach is likely to lead to improved reliability. A summary is included in Appendix I. The interpretation of the structure of motion boundaries has been investigated in human vision. The specific technique developed under a previous AFOSR contract has been found to be used in human vision, representing the discovery of a new perceptual depth cue. Such discoveries in perceptual psychology are both rare and significant. Appendix II includes a reprint giving more information.

Work is continuing integration of motion and static information and on exploiting these results in a variety of image understanding tasks.

Robust methods for determining object motion.

One important function of a vision system is to recognize the presence of moving objects in a scene. If the camera is stationary and illumination constant, this can be done by simple techniques which compare successive image frames, looking for significant differences. If the camera is moving, the problem is considerably more complex. For the purposes of this discussion, *moving objects* are taken to be any objects moving with respect to the stationary portions of the scene, which we refer to as the *environment*. For a moving camera, both moving objects and stationary portions of the scene may be changing position with respect to the camera and thus generating visual motion in the imagery. A moving camera leads to difficulties because of the need to determine objects moving with respect to the environment, rather than the much easier problem of finding objects moving with respect to the camera.

Detection using visual information alone is quite difficult, particularly when the camera is also moving. The availability of additional information about camera motion and/or scene structure greatly simplifies the problem. We develop detection algorithms for the cases in which 1) camera motion is known, 2) only camera rotation is known, 3) only camera translation is known, 4) objects move in contact with a smooth surface, and 5) an object is being actively tracked, but the camera motion associated with the tracking is not known precisely. Appendix III contains a copy of a paper currently under review which describes these results in more detail.

Current optical flow based techniques for estimating parameters of object motion are almost always based on using only local spatial derivatives of optical flow. We have been investigating an alternate approach to solving these problems -- the use of temporal derivatives of flow. The new approach has two advantages. The use of temporal information allows the incorporation of information acquired over a longer time interval, not just a single frame pair. Once temporal derivative based methods are better understood, it will be possible to construct motion analysis algorithms that use both spatial and temporal variation as input. Appendix IV contains the introduction to a paper on this topic that is currently in preparation.

c. Publications.

W.B. Thompson, "Comments on 'Expert' Vision Systems: Some Issues," *Computer Vision, Graphics, and Image Processing,* April 1986.

J.K. Kearney and W.B. Thompson, "Inexact Vision," invited paper, *Proceedings Workshop on Motion: Representation and Analysis,* May 1986.

J.K. Kearney and W.B. Thompson, "Bounding Constraint Propagation for Optical Flow Estimation," in *Motion Understanding: Robot and Human Vision,* W.N. Martin and J.K. Aggarwal, eds., Kluwer Press Academic Publishers, 1986 (with J.K. Kearney).

A. Yonas, L.G. Craton, and W.B. Thompson, "Relative Motion -- Kinetic Information for the Order of Depth at an Edge," *Perception & Psychophysics,* January 1987.

J.K. Kearney and W.B. Thompson, "An error analysis of gradient-based methods for optical flow estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* March 1987.

d. Scientific Collaborators.

Research Assistants:
- Lincoln Craton
- Ian Horswill
- Steven Savitt
- Rand Whillock

Collaborating Faculty:
- Herbert Pick
- Ting-Chuen Pong
- Albert Yonas

Appendix I – Combining motion and non-motion edge cues.

Motion based segmentation uses a combination of traditional segmentation techniques and motion information to yield an improved segmentation of moving objects in a scene. In addition, motion based segmentation provides a 3 dimensional information of the surfaces on either side of the edges found. Motion based segmentation overcomes many of the difficulties of both traditional segmentation techniques and motion techniques while retaining the benefits of both. Motion based segmentation can be implemented efficiently.

Many vision tasks deal with regions as basic image components. Segmentation, which is the process of breaking an image up into regions, is a very important step in these region based tasks. A common method of doing segmentation is to separate the regions by finding the boundaries between the different regions in the image.

There are however basic problems with the way that segmentation is usually done. Segmentation is traditionally based on finding discontinuities certain features of the image such as brightness, color, or texture and then interpreting the strongest discontinuities as edges or boundaries. A major source of problems with this approach is that there are many different causes for discontinuities in an image. Discontinuities can be caused by illumination changes, depth changes, surface markings, or shadows. Of these different causes, only the depth changes give information about the actual structure of objects in the scene. Discontinuities caused by other factors cause false edges which do not correspond to object boundaries. Another disadvantage of traditional techniques is that they do not give any 3 dimensional information about objects in the scene.

Looking at the motion in a sequence of images is also a way of getting information about object boundaries in a scene. A major advantage of using motion information is that all motion boundaries correspond to depth discontinuities which correspond to object boundaries. There are no false edges found. Motion information can also provide 3 dimensional structural information around the boundaries. Using motion information it is possible to determine which side of a boundary is being occluded and which side is occluding.

There are however some problems with using motion information for segmentation. One problem stems from the fact that it is necessary to match points across successive frames to get the motion information. This matching is difficult and any errors in matching causes noise in the motion information which in turn leads to inaccuracies in the edges found between regions. Another problem in using motion information is due to the discreteness of most motion information. Motion information is not usually calculated for every point in the image, but rather for a sampling of points with distinguishing features that will be easy to match. This discrete sampling leads to some inaccuracy in the exact location of the motion information. Which in turn leads to inaccuracies in the position with which edges are located. For example, if motion information is only calculated for one in every ten pixels, the position of any edge is only accurate to within ten pixels. It would be possible to calculate motion information using a larger sample of pixels, this however would greatly increase the computation necessary and would make the matching process more difficult and inaccurate. A decrease in accuracy would result because there would be more possible matches for each sample point and the features between the points would be less distinct.

Motion based segmentation compliments traditional segmentation techniques with motion information. Motion based segmentation is implemented by first running a traditional $\nabla^2 G$ based edge detector to find edges. Motion information is calculated by matching sample points over successive frames. The motion information is then used to filter out the static edges. Edges that have the same motion on both sides are eliminated as false edges, leaving only edges that correspond to actual object boundaries. The final step is to use the motion information to assign a 3 dimensional interpretation at the boundaries. By looking at the motion information around the boarder of an object, and the sign of the $\nabla^2 G$ function, it is possible to determine which side of the edge corresponds to the occluding surface causing the boundary.

The benefits of motion based segmentation are a combination of the benefits from both the traditional segmentation and the motion information. A major benefit is the reduction in false edges. The edges resulting from motion based segmentation correspond only to object boundaries. Another benefit is the positional accuracy of the edges. Since the actual edges are found with the traditional segmenter and the motion information is only used to filter the edges, the position of the edges can be found accurately. Improved accuracy can also be gained through the use of multiple frames. The more frames that points are tracked over the better the motion information. The better the motion information the better the segmentation results. Another and very important advantage of motion based segmentation is the 3 dimensional structural information provided at the boundaries. This structural information can greatly aid classification tasks by associating edges with particular regions. This allows a classifier to work on only those edges that belong to a particular region. Eliminating false edges and edges that belong to other objects in the scene can greatly improve the accuracy of classifiers and other tasks that operate on regions.

Another advantage of the motion based segmentation approach is efficiency. The motion calculation can be done quickly since it only needs to be operate on a sample of points from the scene. The determination of the occluding and occluded surfaces is extremely efficient since it involves only a sign check of the $\nabla^2 G$ function at the points on either side of the boundary.

Appendix II.

Perception & Psychophysics
1987, 41 (1), 53-59

Relative motion: Kinetic information for the order of depth at an edge

ALBERT YONAS, LINCOLN G. CRATON, and WILLIAM B. THOMPSON
University of Minnesota, Minneapolis, Minnesota

A new source of kinetic information for depth at an edge was investigated with adult subjects. The relationship between the motion of optical texture, indicating a surface, and the motion of a contour, indicating an edge, determines whether the surface is perceived as occluding or occluded. Subjects viewed computer-generated random-dot displays in which this relative-motion information provided the only information for depth order and a second type of display in which order in depth was specified both by relative-motion information and by the accretion and deletion of texture. Reliable depth effects were obtained in both conditions. These results indicate that adults are sensitive to the relative motion of texture and contour as information for depth at an edge.

Some depth cues (e.g., binocular disparity) provide the visual system with metric information for spatial layout, specifying the amount of depth separating two objects, whereas other cues (e.g., static interposition) provide only ordinal information. The latter type of cue indicates that a surface is either in front of or behind another, without providing information about the amount of separation in depth. Depth cues may also be classified as static or kinetic. Until recently, most theories of spatial perception have emphasized static depth cues and have paid little attention to the information carried by motion. Gibson (1950) challenged this approach when he pointed out that optical motions resulting from motion of the observer and/or objects provide a rich source of information for detecting the spatial layout of the environment. The purpose of this paper is to describe a potential source of kinetic information for the order of surfaces in depth and to demonstrate that humans are sensitive to this information.

Michotte, Thines, and Crabbe (1964) described a type of display in which purely kinetic information generated the perception of order in depth of two surfaces. Michotte et al. projected a series of forms on a screen, the first of which was a complete circular disk against a dark background. As the sequence continued, more and more of the circle was "blacked out," until it eventually disappeared (see Figure 1). This sequence was also presented in the reverse order. In both cases, the rectilinear portion of the edge of the disk underwent lateral motion while the curved portion appeared immobile. Subjects who viewed these displays reported seeing an unchanging circular disk being covered and uncovered by a second surface. Michotte et al. (1964) referred to this and similar phenomena as the *screening effect*. They maintained that Gestalt laws of perceptual organization accounted for the effect by completing and giving phenomenal permanence to the transforming disk.

Figure 1. Schematic drawing of successive frames from the screening effect display created by Michotte, Thines, and Crabbe (1964).

Both Gibson's (1966, 1979) notion of an ecological optics and Marr's (1982) notion of a computational level for understanding vision argue for an analysis of the structure of the real world and the patterns of proximal stimulation that result from this structure that would make perception of spatial layout possible. Gibson (1966) described the stimulus information in Michotte's displays of the screening effect as the progressive "wiping out" or "unwiping" of optical texture that, he maintained, occurs whenever one surface is covered or uncovered by a second surface. Gibson and his students (Gibson, Kaplan, Reynolds, & Wheeler, 1969) provided a second description of the change in the structure of the optic array that occurs when a surface is covered and uncovered by another:

> When the edge of one surface conceals or reveals a second ... the adjacent units of optical texture on one side of a possible division in the optic array are preserved while adjacent units of optical texture on the other side of the division are progressively added to the array (uncovering) or are progressively subtracted from the array (covering). The decrementing of texture corresponds to a surface being concealed while the incrementing of texture corresponds to a surface being revealed. That side of the dividing line on

This work was supported by NICHD Grant HD-16924 and by Air Force Office of Scientific Research Contract AFOSR-86-0007. The authors wish to thank Herbert Pick for the suggestion that we use a subjective contour display, Kim Pearson and Chi Ping Sze for programming the displays, Martha Arterberry and Herbert Pick for comments on an earlier draft of this manuscript, and Kaye O'Geay for typing the manuscript. Correspondence and requests for reprints should be sent to Albert Yonas, Institute of Child Development, University of Minnesota, 51 East River Road, Minneapolis, MN 55455.

> which there is deletion or accretion always corresponds to the surface that is behind; that side on which there is neither, always corresponds to the surface that is in front. (p. 114)

(It should be noted that the above analysis is not in fact correct for the rotation in depth of smooth surfaces. In such situations, the visible surface of the object is occluding itself. The side of the object boundary on which deletion or accretion is occurring is actually in front of the surface on the other side of the boundary.)

Kaplan (1969) tested two versions of the accretion/deletion hypothesis by presenting adult subjects with an animated film of random texture undergoing progressive accretion and deletion. When these displays are stationary, the viewer perceives a single textured surface. When the texture in these displays undergoes lateral motion, a vertical "subjective contour" is perceived at the vertical margin where texture elements are accreted and deleted. Kaplan presented subjects with three types of accretion/deletion displays. In the first type, texture was accreted or deleted on one side of a stationary vertical subjective contour while texture on the other side of the contour was preserved. This condition tested Gibson's hypothesis that whenever there is accretion/deletion of units of optical texture on one side of the contour and preservation of optical texture on the other side, a depth edge is perceived such that the region undergoing accretion/deletion is seen as a surface that is occluded. The second type of display used by Kaplan tested his own hypothesis that it is the region that undergoes the greater amount of accretion/deletion per interval of time that is perceived as a surface that is occluded. In this condition, texture was accreted or deleted simultaneously on both sides of the subjective contour, at varying relative rates. In a third type of display used by Kaplan, texture was simultaneously accreted or deleted at varying relative rates on both sides of a laterally moving subjective contour. This last condition was critical in that it allowed Kaplan to keep the rate of accretion/deletion constant while varying the velocity and direction of motion of the two surfaces defined by the contour in the display. Kaplan found support for the more general hypothesis that given a difference in the rate of accretion/deletion on the two sides of the subjective contour, adults will perceive depth at an edge.

However, both Kaplan's displays and those used in more recent studies of kinetic occlusion with adults (Andersen & Braunstein, 1983) and infants (Granrud et al., 1984) contain another potential source of information for the covering and uncovering of one surface by another. Thompson, Mutch, and Berzins (1985) have observed that the order of surfaces in depth is specified by the relationship between the optical motion of a contour and the optical motion of the texture elements on either side of the contour. The principle underlying this account is that, for translational motion, the image of an occluding edge moves with the image of the occluding surface to which it belongs. Figure 2 illustrates the effect for simple translational motion. Shown in the figure are the optical motion of texture elements corresponding to two surfaces and the optical motion of a contour corresponding to a depth edge. In Figure 2a, the left surface is in front and occluding the surface to the right. In Figure 2b, the left surface is now behind and being occluded by the surface to the right. The two cases can be distinguished because in Figure 2a the contour moves to the left, whereas in Figure 2b the contour moves to the right.

Although the preceding analysis indicates that relative motion of contour and surface texture is physically determined by the order of surfaces in depth, it is not known whether this information is picked up by the human visual system. In the present study, therefore, we attempted to answer the question: does the relative motion of texture and contour determine the perception of depth at an edge? We investigated this question by presenting subjects with kinetic random-dot displays in which texture on both sides of a centrally located vertical contour was separated from the contour by blank space. This "gap" allowed for lateral motion of texture and contour without providing the subject with accretion/deletion information for depth at an edge. We hypothesized that when a contour and texture on one side of the contour moved together, subjects would perceive a continuous surface that was closer in depth than the surface defined by texture on the other side of the contour, whose motion was not tied to the motion of the contour. The size of the gap between texture and contour was varied to explore whether the detection of relative motion information for depth order was influenced by the spatial separation of the motions. A mechanism that detected relative-motion information might, we thought, function locally and involve processes with relatively small receptive fields. If this were the case, the perception of depth order would become ambiguous as the width of the gap between texture and contour was increased. If the mechanism were more global and integrated information over a large area, the size of the gap should have no effect on the perception of the order of surfaces in depth.

We also investigated whether or not visual mechanisms that detect information for the order of surfaces in depth were able to utilize different sorts of contour information equally effectively. Stimulus displays thus included both objective-contour conditions, in which an ordinary vertical line served as a contour, and subjective-contour conditions. Subjective contours are edges perceived in static displays in the absence of luminance differences (Kanizsa, 1955, 1979; Schumann, 1904). As previously mentioned, subjective contours also occur in conjunction with the depth effect produced by Kaplan's (1969) accretion/deletion displays. We hypothesized that both objective-contour and subjective-contour conditions would produce reliable depth effects.

METHOD

Subjects

Sixteen unpaid students at the University of Minnesota, 15 undergraduates and 1 graduate student, served as subjects.

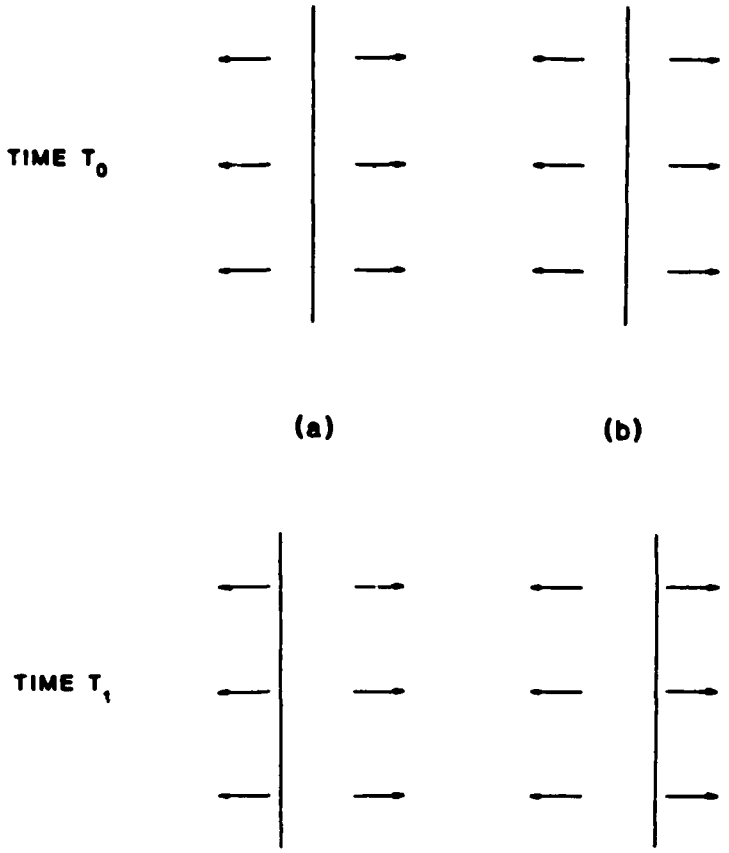


Figure 2. Motion of texture, indicated by arrows, and contours, indicated by change in position of vertical line from T_0 to T_1, specifies in (a) that the surface to the left of the line occludes the surface to the right. In (b), the left surface is occluded by the right surface.

Apparatus

A TERAK microcomputer was used to generate random-dot displays containing kinetic information for the order of surfaces in depth. Displays were presented on a CRT and observed through an 11-cm-high and 21-cm-wide aperture (viewing distance, .91 m; visual angle, 6.9° vertically and 13° horizontally) attached to the face of the CRT. An eyepatch worn over one eye eliminated binocular information. A small box with three buttons labeled "left," "equal," and "right," respectively, was located directly in front of the subject. A 7½-W nightlight provided dim overall illumination for the experimental room, sufficient for the dark-adapted subjects to see the response buttons.

Displays

Eight display conditions were presented. In each of these conditions, a vertical contour was displayed in the center of the screen. Randomly distributed texture elements to the left and right of the contour formed two texture fields. The texture fields contained approximately 2 dots per square centimeter. All the texture elements in a texture field moved in synchronous horizontal motion across the screen at 2 cm/sec. Texture on the left and right traveled simultaneously in opposite directions for a distance of 1 cm and then reversed direction. Thus, the two texture fields alternately approached and receded from one another. This pattern was repeated continuously. On each trial, the lateral motion of the vertical contour in the center of the screen was identical to that of one of the two texture fields. The eight display conditions were as follows:

Display Condition 1. Displays in the accretion/deletion (subjective contour) condition (see Figure 3a) were a modified version of those used by Kaplan (1969). Lateral movements of a subjective contour were accompanied by the appearance and disappearance of individual texture elements, resulting in the progressive accretion and deletion of texture elements on the left side of the display. In addition, the lateral motion of the texture field on the right side of the display was identical to the motion of the contour, whereas the motion of the texture field on the left side was not tied to that of the contour. Thus, in this condition, information for depth was provided by both the accretion and deletion of texture elements and the relative motion of texture and the contour.

Display Condition 2. Displays in the accretion/deletion (objective-contour) condition were identical to those in the accretion/deletion (subjective-contour) condition, except that the central contour was a vertical line rather than a subjective contour. Again, the relative motion of both texture and contour and the accretion and deletion of texture elements provided information for depth.

Display Conditions 3, 4, and 5. In the relative-motion (objective-contour) condition, the relative motion of texture fields and the contour provided the only information for depth (see Figure 3b). Because of the presence of a textureless "gap" between texture fields and contour, this type of display eliminated accretion/deletion information while preserving relative-motion information. Texture on one side (in Figure 3b, the right side) of the display moved with the central contour, so that the gap width between this texture field and contour remained constant. The width of the gap between con-

tour and the texture field not tied to the motion of the contour (in Figure 3b, the texture on the left) changed as the two sides of the display approached and receded from one another. The width of the constant gap was varied to produce gaps of three sizes (1, 3.5, and 5 cm, respectively). For the small-gap, medium-gap, and large-gap conditions, the width of the gap not linked to the contour by relative motion ranged from 0 to 2 cm, from 2.5 to 4.5 cm, and from 4 to 6 cm, respectively. Thus, the average width of the varying gap was approximately equal to the width of the unvarying gap.

Display Condition 6. Displays in the relative-motion (subjective-contour) condition also contained only relative-motion information (see Figure 3c). In the Figure 3c display, a static vertical subjective contour results from the presence of horizontal lines that begin on either side of the display and end at the same vertical midline. Lateral movements of the subjective contour were created in this condition by lengthening the horizontal lines on one side while the horizontal lines on the other side were simultaneously shortened by the same amount. As with the relative-motion (objective-contour) condition, a small gap (1 cm) between texture and contour allowed for movement of the contour without the accretion/deletion of texture elements. Although the changing lengths of the horizontal lines themselves might be interpreted as accretion/deletion, the accretion (lengthening) of one set of lines was always balanced by an identical deletion (shortening) of the lines on the other side of the display. Thus, the accretion/deletion information available in these displays did not specify any depth ordering of surfaces (Gibson et al., 1969; Kaplan, 1969).

Display Conditions 7 and 8. Two types of displays that lacked both relative-motion information and accretion/deletion information served as control conditions. In the first control condition, a single vertical line moved alternately left and right (1 cm in each direction) across a dark homogeneous background. The second control condition was identical to the small-gap relative-motion (objective-contour) condition described above, except that there was no vertical line to serve as a contour. In this case, two laterally moving texture fields alternately approached and receded from each other. The amount of separation between the two texture fields in this condition ranged from 1 to 3 cm.

Procedure

Before being tested, the subjects were given the following instructions: "Look at the display and decide which of the following is true: (1) the left side looks like it is in front; (2) the right side looks like it is in front; (3) both sides appear to be the same distance away."

The subjects were instructed to record their responses by pressing one of the three buttons, labeled "left," "equal," and "right," on the small box directly in front of them. Each subject received three practice trials in which two sheets of paper were held in front of the CRT by the experimenter and moved laterally so that they approached and receded from each other in a manner analogous to the computer-generated displays described above. The relative depth of the sheets of paper was varied so that the one on the left was closer to the subject on one trial, the one on the right was closer to the subject on another trial, and the sheets were equidistant from the subject on a third trial. The practice trials were presented in random order, and the subjects were asked to indicate which button represented the appropriate response. All subjects performed quickly and without error on the practice trials.

During test trials, the subjects observed displays monocularly from a distance of 3 ft (.91 m). The subjects viewed eight blocks of trials, corresponding to the eight conditions described above, with 20 trials in each block. The order of presentation for blocks was completely counterbalanced for the group as a whole. A new random distribution of texture elements was generated for each block of trials.

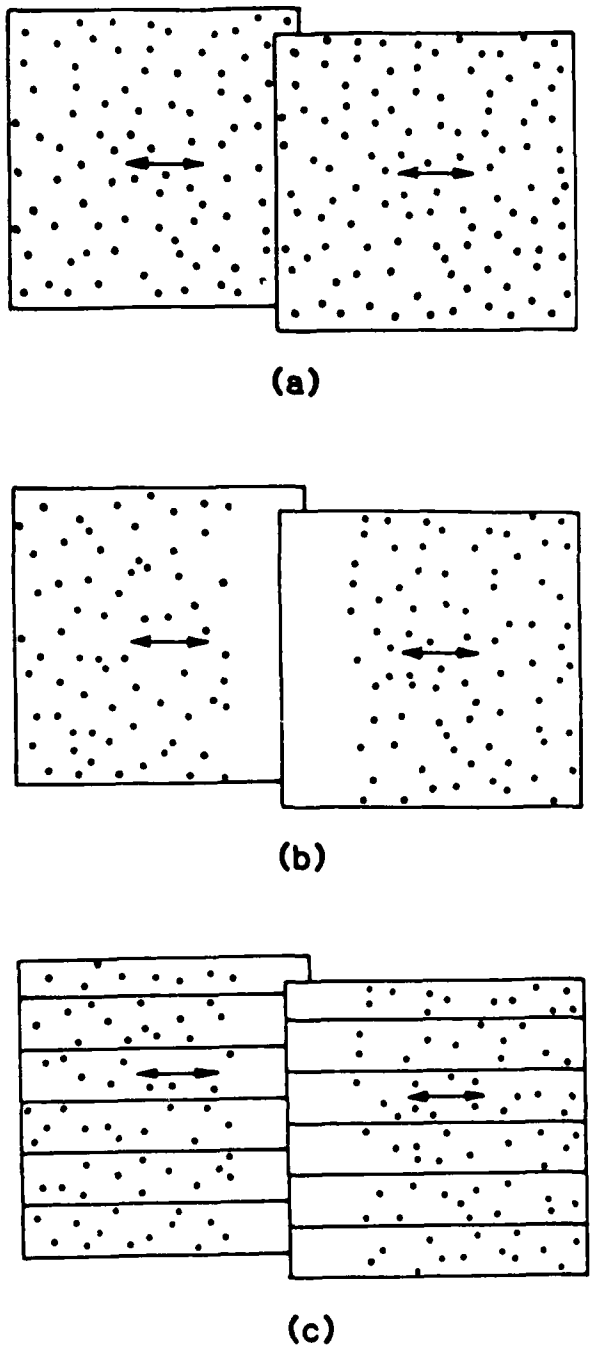


Figure 3. (a) Schematic drawing of accretion/deletion displays. In the subjective-contour condition, a subjective contour is perceived at the margin where accretion and deletion of texture occurs. In the objective-contour condition, a vertical line is located at the margin. (b) Relative-motion (subjective-contour) display. "Gap" between vertical line and texture eliminates accretion/deletion information. (c) Relative-motion (subjective-contour) display. End-stopped horizontal lines generate a vertical subjective contour. Interposition displayed in these drawings was not present in computer-generated displays.

Within a given block, the same random sequence of predicted depth orderings (right in front, left in front, neither in front) was presented to each subject. All displays continued without interruption until the subject recorded a response. After each response, there was a brief pause and then the next display was initiated. Total testing time was approximately 25 min.

RESULTS AND DISCUSSION

The mean percentage of depth judgments consistent with predicted depth order and opposite predicted depth order and the mean percentage of "no depth" responses are presented in Table 1. Two mixed-model repeated measures analyses of variance (ANOVAs) were conducted. To determine whether the experimental conditions yielded the perception of depth at an edge, a one-way ANOVA was carried out on the mean number of "no depth" responses for each of the eight conditions. This analysis revealed a significant main effect for condition [$F(7,105) = 28.07$, $p < .01$]. Post hoc comparisons based on Tukey's honestly significant difference method indicated that all six experimental conditions yielded significantly fewer responses of "no depth" than either control condition ($p < .05$). To establish whether there were differences between the six experimental conditions in determining the perceived order of surfaces in depth, a second one-way ANOVA was carried out on the mean number of responses consistent with predicted depth order. This analysis yielded a significant main effect for condition [$F(5,75) = 10.30$, $p < .05$]. Tukey post hoc comparisons revealed several significant differences ($p < .05$), as shown in Table 2.

As the data in Table 1 show, the modified version of Kaplan's (1969) accretion/deletion (subjective-contour) displays used in this study were effective in producing the perception of depth at an edge. In addition, the reliable depth effect obtained in the accretion/deletion (objective-contour) condition indicates that visual processes sensitive to the depth information in these accretion/deletion displays are not disrupted when an ordinary vertical line serves as a contour.

Table 1
Mean Percentage of Depth Order Judgments as a Function of Display Condition

	Predicted Depth Order		No Depth		Opposite Predicted Depth Order	
Condition	Mean	*SD*	Mean	*SD*	Mean	*SD*
Accretion/deletion						
Subjective contour	98.8	2.8	0.6	0.5	0.6	1.6
Objective contour	98.1	7.3	0		1.9	7.3
Relative motion						
Subjective contour						
Small gap	95.6	9.2	3.1	6.6	1.3	1.2
Objective contour						
Small gap	87.2	16.7	10.0	15.8	2.8	4.7
Medium gap	78.8	19.7	16.5	16.7	4.7	7.0
Large gap	62.5	33.7	24.4	27.5	13.1	18.1
Controls						
Contour only			67.5	30.7		
Texture only			53.5	31.4		

The main purpose of the present study was to determine whether adults were sensitive to depth information specified by the relationship between the optical motion of a contour and the optical motion of texture elements on both sides of the contour. The depth effect obtained in each of the relative-motion conditions (see Table 1) indicates that when depth information from the accretion and deletion of texture is eliminated, adults perceive depth at an edge on the basis of relative-motion information alone. In addition, as the data in Table 2 show, the two relative-motion conditions with small gaps (the subjective-contour condition and the objective-contour/small-gap condition) did not differ significantly from either of the two accretion/deletion conditions in determining the perceived order of depth.

It is possible, however, that subjects' judgments in the relative-motion conditions were based upon a response bias produced by the practice trials and/or the accretion/deletion displays. Prior exposure to these may have created a set, or response criterion, for interpreting the relative-motion displays. To rule out this possibility, a follow-up experiment was conducted in which 7 naive adult subjects viewed only a single continuous display of the small-gap relative-motion (objective-contour) condition. The procedure employed was similar to that used in the main study, except that the practice trials were eliminated and the subjects were simply asked to describe the display. Initially, no mention of depth was made. Three of the 7 subjects spontaneously reported seeing one surface moving over another, with the predicted depth order. When prompted with the question "Is there any depth suggested in the display?" the remaining 4 subjects all reported the predicted depth effect. This was the case even though no mention was made of what form the apparent depth might take. Thus, the depth effect obtained from relative-motion information in the absence of other cues seems quite robust; to our knowledge, sensitivity to this depth cue has not been previously demonstrated.

A second finding of the study may be informative about the mechanism used by the visual system in perceiving depth from relative-motion information. As the data in Table 2 show, both of the small-gap relative-motion (subjective-contour and objective-contour) conditions yielded significantly more responses consistent with predicted depth order than did the large-gap relative-motion (objective-contour) condition. In addition, as Table 1 indicates, the large-gap condition showed significantly more responses of "no depth" than did either of the accretion/deletion conditions. These results indicate that the depth effect obtained in the relative-motion conditions diminishes as the width of the gap between texture and contour is increased. One interpretation of this finding is that the detection of relative-motion information for depth depends on processes that are relatively local, namely, computations that compare the motion of texture elements that are relatively near the contour with the motion of the contour. However, it should be noted that even the large-gap relative-motion condition produced a perception of depth relative to control conditions (see Ta-

Appendix III.

DETECTING MOVING OBJECTS

William B. Thompson
Ting-Chuen Pong

Computer Science Department
University of Minnesota
Minneapolis, MN 55455

ABSTRACT

The detection of moving objects is important in many tasks. This paper examines moving object detection based primarily on visual motion. We conclude that in realistic situations, detection using visual information alone is quite difficult, particularly when the camera is also moving. The availability of additional information about camera motion and/or scene structure greatly simplifies the problem. We develop detection algorithms for the cases in which 1) camera motion is known, 2) only camera rotation is known, 3) only camera translation is known, 4) objects move in contact with a smooth surface, and 5) an object is being actively tracked, but the camera motion associated with the tracking is not known precisely. Examples of several of these techniques are presented.

1. Introduction.

One important function of a vision system is to recognize the presence of moving objects in a scene. If the camera is stationary and illumination constant, this can be done by simple techniques which compare successive image frames, looking for significant differences. If the camera is moving, the problem is considerably more complex. For the purposes of this discussion, *moving objects* are taken to be any objects moving with respect to the stationary portions of the scene, which we refer to as the *environment*. For a moving camera, both moving objects and stationary portions of the scene may be changing position with respect to the camera and thus generating visual motion in the imagery. A moving camera leads to difficulties because of the need to determine objects moving with respect to the environment, rather than the much easier problem of finding objects moving with respect to the camera. In this paper, we deal with the problem of detecting moving objects from a moving camera based on optical flow

The visual detection of moving objects is a surprisingly difficult task. A simple example illustrates just how serious the problem can be. Consider the optical flow field shown in

This work was supported by Air Force Office of Scientific Research contract AFOSR-85-0382

figure 1, which appears to show a small, square region in the center of the image moving to the right and surrounded by an apparently stationary background. Such a flow field can arise from several equally plausible situations: 1) The camera is stationary with respect to the environment, and the central region corresponds to an object moving to the right. 2) The camera is moving to the left with respect to the environment, most of the environment is sufficiently distant so that the generated optical flow is effectively zero, while the central region corresponds to a surface near to the camera but stationary with respect to the environment. 3) The camera and object are moving with respect to both the environment and each other, though the environment is sufficiently distant so that there is no perceived optical flow. It is not possible to tell whether or not this seemingly simple pattern corresponds to a moving object!

Figure 1 provides one example of why a general and reliable solution to the problem of moving object detection based only on visual motion is not feasible. Robust solutions require that additional information about camera motion and/or scene structure be available. In this paper, we examine a variety of types of information that might be available. Each information source places constraints on the optical flow fields that can be generated by a camera moving through an otherwise static environment. Violations of these constraints are thus necessarily due to moving objects.

Figure 2 summarizes potential sources of information and the associated constraints on optical flow. The next section lists general properties needed by reliable detection algorithms. Following this is a derivation of each of the flow constraints. We conclude with experimental demonstration of several of the techniques and general observations about the nature of these methods.

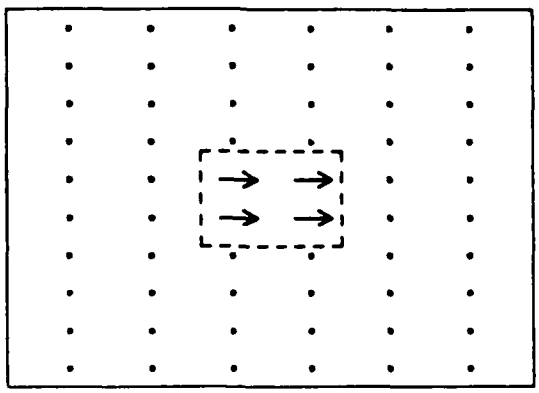


Figure 1: Is The Central Region a Moving Object?

Knowing:	Yields a constraint on:
full parameters of motion	flow values
parameters of rotation	variability of flow direction
parameters of translation	direction in difference field
surfaces are smooth	local variability of direction or magnitude of flow
object is being tracked	global variability of direction of flow

Figure 2: Constraints on Flow.

2. Assumptions.

We start with the presumption that motion detection algorithms should be designed with the following properties in mind:

The field of view may be relatively narrow.

Motion detection should not depend on the use of wide angle imaging systems. Such systems may not be available in a particular situation, and if used may increase the difficulty or recognizing small moving objects. As a result, detection algorithms should not depend on subtle properties of perspective.

The image of moving objects may be small with respect to the field of view.

This is clearly desirable for reliability. Moving objects may be far away and subtended by relatively small visual angles. We need methods capable of identifying single image points, or at least small collections of points, as corresponding to moving objects. Detection algorithms thus cannot depend on variations in flow over a potentially moving object.

Only monocular imagery is available.

This is equivalent to the situation where objects of interest can be far away relative to the camera base-line in a stereo viewing situation.

Estimated optical flow fields will be noisy.

No method is capable of estimating optical flow with arbitrary accuracy. Motion detection based on optical flow must be tolerant of noisy input.

Only "instantaneous" optical flow is used.

A restriction to instantaneous flow eliminates the use of temporal derivatives of flow and/or multiple views at distinct time intervals. Temporal differentiation will increase noise in the estimated flow values. Use of multiple views increase computational complexity. (In fact, experience with There are reasons to believe that multi-frame analysis techniques may in fact improve reliability [1], though they are not examined in this work.)

3. Constraints on Optical Flow.

The basic mathematics governing the optical flow generated by a moving camera is well known. We take our notation from [2], using a coordinate system fixed to the camera (e.g. the world can be thought of as moving by a stationary camera). Optical flow values are a function of image location, the relative motion between the camera and the surface point corresponding to the image location, and the distance from the camera to the corresponding surface point. Let $\mathbf{p} = (x, y)$ refer to an image location, where x and y have been normalized by the focal length of the camera. Let $\mathbf{P} = (X, Y, Z)$ be the coordinates of the surface point projecting onto (x, y), specified in a coordinate system with origin at the camera and Z axis along the optical axis of the camera. Specify the motion of the point at (X, Y, Z) with respect to the camera in terms of a translational velocity $\mathbf{T} = (U, V, W)^T$ and a rotational velocity $\omega = (A, B, C)^T$. The optical flow, $\mathbf{B} = (u, v)$, at $\mathbf{p}$ is purely a function of x, y, $\mathbf{T}$, ω, and Z:

$$u = u_t + u_r \quad , \quad v = v_t + v_r \tag{1}$$

where u is the x component of flow, v is the y component of flow, and

$$u_t = \frac{-U + xW}{Z} \quad , \quad v_t = \frac{-V + yW}{Z} \tag{2}$$

$$u_r = Axy - B(x^2 + 1) + Cy \quad , \quad v_r = A(y^2 + 1) - Bxy - Cx \tag{3}$$

Let the parameters specifying camera motion with respect to the environment be $\mathbf{T}_e$ and ω_e and the corresponding parameters specifying relative motion between the camera and a scene point $\mathbf{P}$ be $\mathbf{T}_\mathbf{P}$ and $\omega_\mathbf{P}$.

3.1. Known translation and rotation.

The parameters of camera motion constrain possible optical flow values that can occur due to camera motion with respect to the environment.

Detecting Moving Objects

If complete information about instantaneous camera motion is available, then $\mathbf{T}_e$ and ω_e are known. If the camera is translating but not rotating with respect to the background, $\omega_e = 0$ and all flow vectors due to the moving image of the background will radiate away from a *focus of expansion* (FOE). From equation (1), it is easy to see that the image plane location of the FOE is at:

$$x_{foe} = \frac{U}{W} \quad , \quad y_{foe} = \frac{V}{W} \quad . \tag{4}$$

The location of the FOE depends only on the direction of translation, not on the speed, so methods for motion detection which depend on the location of the FOE do not actually require the complete parameters of translational motion. The FOE may not lie within the visible portion of the image (and in fact may be a focus of contraction). A FOE at ∞ corresponds to pure lateral motion, which generates a parallel optical flow pattern. At every image point **p**, knowing the FOE fully specifies the direction of optical flow associated with any surface point stationary with respect to the environment. At **p**:

$$\theta_{foe} = \tan^{-1} \frac{V - Wy}{U - Wx} \quad , \quad \theta_{flow} = \tan^{-1} \frac{v_t}{u_t} \tag{5}$$

where θ_{foe} is the direction from **p** towards the FOE and θ_{flow} is the direction of optical flow at **p**. (Note that the first equation is still well defined even if $W = 0$, corresponding to a focus of expansion at ∞ in image coordinates.) Any flow values with a different direction correspond to moving objects [3]. E.g., moving objects exist whenever $\|\theta_{foe} - \theta_{flow}\| > \epsilon$, for some appropriate ϵ. (It is possible that moving objects coincidentally generate flow values compatible with this constraint.) This approach requires the estimation of only the direction of flow, not either the magnitude or spatial variation of flow.

Camera rotation introduces considerable complexity. Knowledge of camera motion no longer constrains the direction of background flow. Nevertheless, at a given point **p**, flow is constrainted to a one-dimensional family of possible vector values. The family is given by (1) – (3) where Z ranges over all positive values. The analysis can be simplified because of the linear nature of (1). u_r and v_r depend only on the parameters of rotation and not on any shape property of the environment. Because the values of u_r and v_r at a particular point **p** do not depend on Z, they can be predicted knowing only ω. These values can be subtracted from the observed optical flow field, leaving a *translational flow field*:

$$\mathbf{F}_t = (u_t, v_t) = \mathbf{F} - \mathbf{F}_r \quad , \quad \mathbf{F}_r = (u_r, v_r) \tag{6}$$

where u_r and v_r are defined in equation (3). This field behaves just as if no rotation was occurring, and thus moving objects can be located using the FOE technique described above. For the remainder of this paper, when rotation is present, we will take the term FOE to refer to the focus of expansion of this translational field.

In principle, even if camera motion is not known $\mathbf{T}_e$ and ω_e may be estimated from the imagery [2], subject to a positive, multiplicative scale factor for $\mathbf{T}_e$. Two serious problems exist, however. Narrow angles of view make estimation of camera motion difficult, as significantly different parameters of motion and surface shape can yield nearly identical optical flow patters [4]. In addition, techniques such as [2] uses a global minimization approach which will not perform well if moving objects make up a substantial portion of the field of view. A clustering approach (e.g. [5]) can be made tolerant of the moving objects, great

difficulty can be expected dealing with a five dimensional cluster space.

3.2. Known rotation.

The parameters of camera rotation constrain the local variability of optical flow direction that can occur due to camera motion with respect to the environment.

Non-visual information about camera motion often comes from inertial sources. Such sources are much more accurate in determining rotation than translation. Rotation involves a continuous acceleration which is easily measured. The determination of translation requires the integration of accelerations, along with a starting boundary value. Errors in estimated translation values rapidly accumulate. A simple technique allows the detection of moving objects when only camera rotation is known.

In the previous sections, knowledge of camera rotation made it possible to compute the translational flow field, $\mathbf{F}_t$. Knowledge of translation was then used to locate the FOE and thus constraint the direction of flow vectors associated with the environment. If only rotation is known, then it is still possible to determine the translational flow field, but not the FOE. Visual methods could be applied to the translational flow field to estimate the location of the FOE, but these methods suffer from a number of practical limitations when applied to noisy data. An alternate approach can be used which does not require the prior determination of the FOE. The translational flow field extends radially from the focus of expansion. At any point significantly away from the FOE, the direction of flow (but not necessarily the magnitude of flow) will vary slowly. Directional variability can be evaluated based on equation (5):

$$\frac{\delta \theta_{foe}}{\delta x} = \frac{W(V - yW)}{(V-yW)^2 + (U-xW)^2}, \quad \frac{\delta \theta_{foe}}{\delta y} = -\frac{W(U - xW)}{(V-yW)^2 + (U-xW)^2} \tag{7}$$

The gradient of the direction of the translational flow field can thus be obtained as

$$\left(\frac{\delta \theta_{foe}}{\delta x}\right)^2 + \left(\frac{\delta \theta_{foe}}{\delta y}\right)^2 = \frac{1}{(y_{foe} - y)^2 + (x_{foe} - x)^2} \tag{8}$$

where (x_{foe}, y_{foe}) is the image plane location of the FOE. We can see from the above equation that over any local area away from the FOE, variations in the *direction* of the translational flow field will be small. Flow arising due to moving objects is of course not subject to this restriction. The gradient of flow field direction can thus be used to detect the boundaries of moving objects. At these boundaries, flow direction will vary discontinuously[1].

3.3. Known translation.

The parameters of camera translation constrain the direction of vectors in the "difference field" that can occur due to camera motion with respect to the environment.

Under some circumstances, the trajectory of the camera platform may be known, but

[1] Marr [6] claims "if direction of [visual motion] is ever discontinuous at more than one point – along a line, for example, – then an object boundary is present." Note that this is only necessarily true if no camera rotation is occurring (or equivalently, if camera rotation has been normalized by using the translational flow field)

the camera is undergoing unknown rotations.[2] Because rotation is not known, it is not possible directly normalize for the effects of rotation by computing the translational flow field. Instead, a local differencing technique can be used to eliminate the effects of rotation [7,8]. Large, local changes in flow can occur only due to significant depth discontinuities or due to the presence of surfaces moving with respect to one another. To select flow boundaries actually corresponding to moving objects, a technique similar to the FOE approach can be used. Let **DF** be a *difference field* associated with an optical flow field **F**:

$$\mathbf{DF}(x,y) = \mathbf{F}(x,y) - \mathbf{F}(x-\delta x, y-\delta y) \tag{9}$$

$$= \Big((u_t(x,y)-u_t(x-\delta x,y-\delta y)) + (u_r(x,y)-u_r(x-\delta x,y-\delta y)),\ (v_t(x,y)-v_t(x-\delta x,y-\delta y)) + (v_r(x,y)-v_r(x-\delta x,y-\delta y))\Big) \tag{10}$$

For small δx and δy, the magnitude of **DF** can only be large if either there is a significant change in depth over the interval $(\delta x, \delta y)$ or if the interval spans the boundary of a moving object. If (x,y) and $(x-\delta x, y-\delta y)$ both correspond to locations in the environment, δx and δy are both small, and Z changes significantly over the interval, then:

$$\mathbf{DF}(x,y) \approx \Big((u_t(x,y) - u_t(x-\delta x,y-\delta y)), (v_t(x,y) - v_t(x-\delta x,y-\delta y))\Big) \tag{11}$$

$$\approx \left(1 - \frac{Z(x,y)}{Z(x-\delta x,y-\delta y)}\right)\mathbf{F}_t(x,y) \tag{12}$$

The Z values in the above equation are scalars. As a result, if the interval over which the difference is taken does not span the boundary of a moving object, the value of **DF** is a vector parallel to the corresponding value of $\mathbf{F}_t$, that is it is a vector pointing towards or away from the FOE. If the magnitude of **DF** is large and the direction is not compatible with the FOE, then the a moving object must be present.

[8] suggests using this effect to actually locate the FOE. For a variety of reasons, this may be quite difficult in practice. If camera translation is known, however, the **DF** field may be used to detect moving objects even in the presence of rotation. Large magnitude elements of **DF** are examined. The directions of such elements are then checked for compatibility with the FOE. Incompatible elements correspond to the edges of moving regions. Note that a constraint on scene structure as well as information about camera motion is required. In particular, the method is only effective if there are significant depth discontinuities over visual portions of the environment.

3.4. Motion over smooth surfaces.

Object motion over smooth surfaces constrains the local variability of flow.

Knowledge of the shape of environmental surfaces can be used to simplify the motion detection problem. Scene structure may be known precisely (e.g. the range to visible surface points) or in terms of general properities (e.g. significant depth discontinuities can be expected). Information about scene structure can come from visual sources (e.g stereo [9,10]).

[2] We can expect these situations to be rare. If the direction of translation were known over some interval of time, it would be an easy matter to determine the rotation by examining the rate of change of direction

or from pre-existing models of the environment. If both the optical flow, (u, v), and the depth, Z, are known for a collection of surface points in the environment, then (1) – (3) can be used to create a system of equations which can be solved for the parameters of motion T and ω. If the collection of points includes some values associated with the environment and others associated with one or more objects moving with respect to the environment, the system of equations used to solve for T and ω will be inconsistent. Checking the system for consistency can therefore be used as a test for the presence of a moving object (e.g. a test for non-rigid motion in the field of view.)

If moving objects must remain in contact with environmental surfaces (e.g. vehicular motion), a less complex technique depending only on knowing the image plane locations corresponding to discontinuities in range is possible. If no objects are moving within the field of view, equations (1) – (3) can be simplified into the following form:

$$flow(\mathbf{p}) = f_r(\mathbf{p}) + \frac{f_t(\mathbf{p})}{r(\mathbf{p})} \tag{13}$$

where at an image point p, $flow(\mathbf{p})$ is the optical flow (a two-dimensional vector), f_r is the component of the flow due to the rotation of the scene with respect to the sensor, f_t is dependent on the translational motion of the sensor and the viewing angle relative to the direction of translation, and r is the distance between the sensor and the surface visible at p (i.e. the value of Z in equation corresponding to the image location p). For fixed p, flow varies inversely with distance. Both f_r and f_t vary slowly (and continuously) with p. Discontinuities in *flow* thus correspond to discontinuities in r. This relationship holds only for relative motion between the camera and a single, rigid structure. When multiple moving objects are present, equation (13) must be modified so that there is a separate $f_r^{(i)}$ and $f_t^{(i)}$ specifying the relative motion between the sensor and each rigid object. Discontinuities in flow can now arise either due to a discontinuity in range or due to the boundaries of a moving object. If independent information is available on the location of range discontinuities, and other discontinuities in flow must be due to moving objects.

The motion detection problem becomes particularly simple if the environment is planar. In this case, depth discontinuities are not possible and *any* discontinuity in flow (either direction or magnitude) corresponds to the boundary of a moving object. Note that it is not sufficient to know simply that the environment is a "smooth" surface. From some viewing positions, even smooth surfaces may exhibit range discontinuities.

3.5. Tracking regions of interest.

Tracking an object constrains the global variability of the direction of flow in the surrounding area.

A vision system which can actively control camera direction is capable of tracking regions of interest over time, keeping some particular object centered within the field of view. Tracking regions of interest is desirable for many reasons other than the detection of moving objects (e.g. [11]), though the analysis of imagery arising from a tracking camera has not received much study by the computer vision community. If there are significant variations in depth over the visible portion of the background and if moving objects are relatively small with respect to the field of view, then moving object detection based on tracking can be accomplished without any actual knowledge of camera motion. (For motion detection,

the tracking can easily be simulated if the camera is not actively controllable.)

If an object is being tracked, then its optical flow is zero. Flow based methods for determining whether or not a tracked object is moving must depend wholly on the patterns of flow in the background. Object tracking helps in moving object detection because it minimizes many of the difficulties due to rotation. When dealing with instantaneous flow fields, we can decompose the problem by considering all translational motion to be due to movement of the camera platform and all rotational motion due to pan and tilt of the camera to accomplish the tracking. (We will disregard any effects due to spin around the line of sight.) Consider the effect of tracking a point that is in fact part of the environment. The translational component of motion induces an optical flow pattern field extends radially from the focus of expansion, with magnitudes dependent on the range to the corresponding surface points. Over a local area away from the focus of expansion, the *direction* of translational flow will be approximately constant. The rotational component of motion induces a flow pattern which over a local area is approximately constant in both direction and magnitude. The magnitude and direction are exactly opposite the translational flow of the tracked point. From equation (02) and (03), it is easy to see that at the tracked point $(x,y) = (0,0)$

$$u_t = -\frac{U}{Z}, \quad v_t = -\frac{V}{Z} \tag{14}$$

$$u_r = -B, \quad v_r = A \tag{15}$$

Since the optical flow is zero at the tracked point, we have

$$-\frac{U}{Z} - B = 0, \quad or \quad u_t = -u_r \tag{16}$$

$$-\frac{V}{Z} + A = 0, \quad or \quad v_t = -v_r \tag{17}$$

The effect on the combined fields is that in the neighborhood of the tracked point, the direction of flow will be approximately constant (modulo 180°), with a magnitude dependent on the difference between the range to the corresponding surface point and the range to the tracked point. Now, consider tracking a point that is moving with respect to the environment. If environmental surface points are visible in the neighborhood of the tracked point, and if there is a variation in range to these environmental points, then there will be a variation in direction of flow over the neighborhood.

4. Examples.

A set of experiments on moving object detection based on the techniques discussed in the previous sections have been preformed on real images. Experimental results are presented in this section for the cases in which 1) the camera rotation is known, 2) objects move in a smooth environment, and 3) a potentially moving object is being actively tracked.

Figure 3 a) and b) show a pair of images of an indoor scene. In this example, the camera rotates and translates with respect to the environment while the toy vehicle on the table moves to the right between image frame 1 and 2. The rotational velocity of the camera with respect to the environment was measured. The optical flow field shown in figure 4 was obtained by the token matching technique described in [12]. The translational flow field shown in figure 5 was obtained by subtracting the rotational flow component computed from

the known rotational velocity from the observed optical flow field (figure 4). The gradient of flow direction in the translational flow field was used to detect the boundaries of moving objects. Figure 6 shows the detected boundary of a moving object overlaid onto the first frame of figure 3.

The pair of images in figure 7 are used to illustrate the technique for detecting objects moving in a smooth environment. In this example, the camera moves with respect to an environment consisting of nuts and bolts lying on a planar surface. The optical flow field shown in figure 8 was obtained in the same manner as in figure 4. Locations corresponding to large variations in optical flow values is considered to be the boundary of a moving object. Figure 9 shows the locations of large variations in optical flow values, corresponding to the boundary of a moving object.

In figure 10, the circular object in the center of the image is being tracked by the camera while the camera is translating to the right with respect to the environment. Figure 11 shows the estimated optical flow. Figure 12 shows a histogram of the directions of the optical flow. Note that there are two distinct peaks in the histogram. The highest peak corresponds to the optical flow vectors associated with the background and the second peak corresponds to the optical flow vectors associated with the box and the table in the foreground. The variation in flow direction over the image was computed to be approximately 26°, indicating that the tracked object was in fact moving.

As a comparison, a similar experiment in which the tracked object is stationary with respect to the environment while the camera is moving was also preformed. A pair of images similar to that of figure 10 were obtained. The resulting estimated optical flow field is shown in Figure 13. Its corresponding histogram is shown in figure 14. Note that only one distinct peak is observed in this histogram. The global variation in flow direction in this case was computed to be approximately 14° which is significantly smaller than that of the previous example.

5. Discussion.

The methods described above can be grouped into three classes. *Point-based* techniques (known motion, known translation) compare individual optical flow vectors against some standard to determine incompatibilities with the motion of the camera relative to the environment. In all cases described here, the compatibility measure is based on a directional constraint associated with the focus of expansion of the translational flow field. Point-based methods have the advantages of computational simplicity and the ability to detect very small moving objects. They will be most effective when parameters of motion are known precisely and the magnitude of the translational flow field at the point in question is sufficiently large to allow an accurate estimate of direction. *Edge-based* techniques (known rotation, smooth surface) roughly correspond to traditional edge detection. Edge-based motion detection is characterized by the differential flow properties examined and by the filtering technique used to separate edges due to range discontinuities from those due to moving objects. The approach is effective when surfaces are smooth and techniques exist for accurately locating those range discontinuities that do exist. Edge-based methods have the advantage of specifying the outline of moving objects that are detected. They are likely to be of limited use when moving objects are quite small. *Region-based* techniques (tracked object) examine optical flow values over a region, searching for distributions incompatible with rigid motion.

As with edge-based approaches, the viewed region must include portions of both object and environment. As long as the region includes portions of both object and environment, this is an effective test for moving objects that does not require any information about camera motion. The region-based method based on tracking potentially moving objects does not require any information about camera motion, but does require that there be significant variations in range over the visible portions of the environment.

One region-based technique not discussed above is based on an explicit check for rigidity. Several structure-from-motion algorithms provide an estimate of rigidity [13,14,15]. Such checks can presumably be used to recognize non-rigid motion due to the presence of a moving object. Numerical structure-from-motion algorithms have proven to be unsatisfactory in practice due to severe problems with ill-conditioning. It is not yet clear whether or not the test for rigidity can be performed in a sufficiently noise tolerant manner to provide for reliable moving object detection.

No method for detecting moving objects will be effective if it depends on knowing precise values of optical flow. Techniques for estimating optical flow are intrinsically noisy (e.g. see [16]). Additional difficulties arise due to the idealized nature of equations (1) - (3). Real cameras are not point projection systems. Substantial effort is required to accurately determine the values of x and y in (2) and (3). Geometric distortions in the optical and sensing systems affect measured locations on the image plane. Variabilities in effective focal length to to focus can be substantial. Reliable techniques will be based on searching for large magnitude effects in the flow field [17]. All of the methods described above compare flow vectors to some predetermined standard, or look for significant differences across flow boundaries. As a result, all deal with relatively large magnitude effects, though reliability is dependent on scene structure, the nature of camera motion, and position in the visual field relative to the direction of translation.

Many of the techniques described above are based on comparing flow values at different points within the field of view. All of these methods require that measurable optical flow exist for points both in the environment and on moving objects. (Some require only that the translational flow be measurable.) Such methods share three important limitations: 1) they are ineffectual near the FOE, 2) the camera must be moving, and 3) portions of the visible environment must be sufficiently close to generate recognizably non-zero translational flow values. Near the FOE, flow due to the environment will be close to zero, regardless of range. If the camera is not moving, all environmental flow values will be zero. The same is true if all points in the environment are very distant relative to the speed of translation. These limitations do not apply just to the methods listed above, as illustrated by figure 1, they are general problems associated with any vision-based motion detection scheme that does not have accurate information about camera translation and/or range to visible surface points.

BIBLIOGRAPHY

[1] T.J. Broida and R. Chellappa, "Estimation of object motion parameters from noisy images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, January 1986.

[2] A.R. Bruss and B.K.P. Horn, "Passive Navigation," *Computer Vision, Graphics, and Image Processing*, v. 21, n. 1, pp. 3-20, 1983.

[3] R.C. Jain, "Segmentation of Frame Sequences Obtained by a Moving Observer," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 5, pp. 624-629. September 1984.

[4] G Adiv, "Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 1985.

[5] D.H. Ballard and O.A. Kimball, "Rigid body motion from depth and optical flow," *Computer Vision, Graphics, and Image Processing*, vol. 21, pp. 3-20, 1983.

[6] D.A. Marr, *Vision*, San Francisco: W.H. Freeman and Company, 1982.

[7] H.C. Longuet-Higgins and K. Prazdny, "The interpretation of a moving retinal image," *Proc. R. Soc. Lond.*, v. B 208, pp. 385-397, 1980.

[8] J.H. Reiger and D.T. Lawton, "Sensor motion and relative depth from difference fields of optic flows," *Proc. 8th Int. Joint Conf. on Artificial Intelligence*, pp. 1027-1031, August 1983.

[9] A.M. Waxman and J.H. Duncan, "Binocular image flows," *Proc. Workshop on Motion Representation and Analysis*, pp. 31-38, May 1986.

[10] T.S. Huang, S.D. Blostein, A. Werkheiser, M. McDonnel, and M. Lew, "Motion detection and estimation from stereo image sequences: Some preliminary experimental results," *Proceedings Workshop on Motion: Representation and Analysis*, May 1986.

[11] A. Bandopadhay, B. Chandra, and D.H. Ballard, "Active navigation: Tracking an environmental point considered beneficial," *Proceedings Workshop on Motion Representation and Analysis*, May 1986.

[12] S.T. Barnard and W.B. Thompson, "Disparity analysis of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-2, pp 333-340, July 1980.

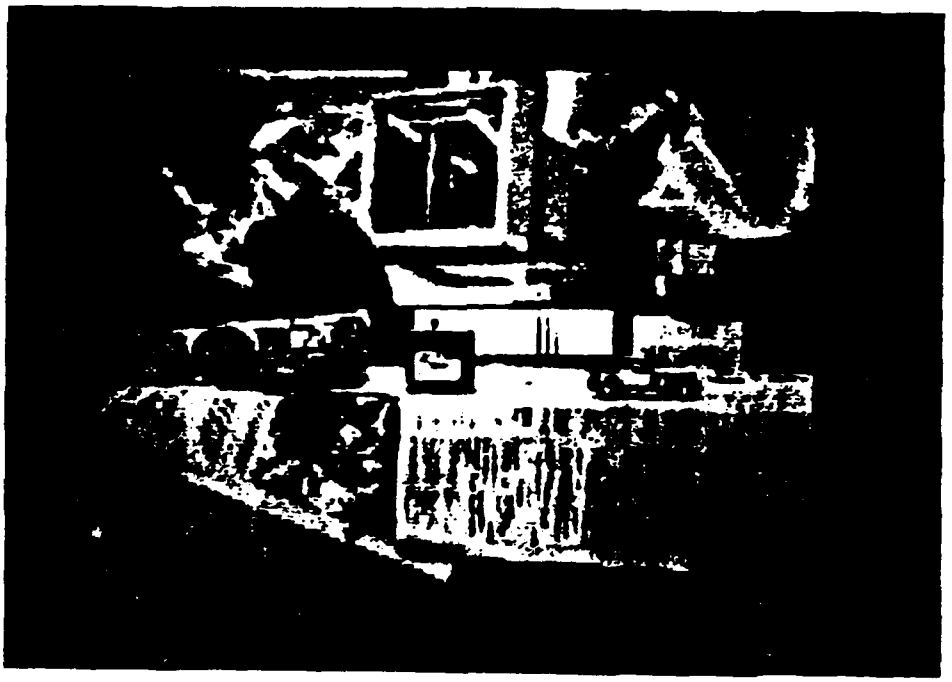
[13] S. Ullman, "Maximizing rigidity: The incremental recovery of 3-D structure from rigid and rubbery motion," *Perception*, 1984.

[14] A. Mitche, S. Seida, and J.K. Aggarwal, "Determining position and displacement in space from images," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 1985.

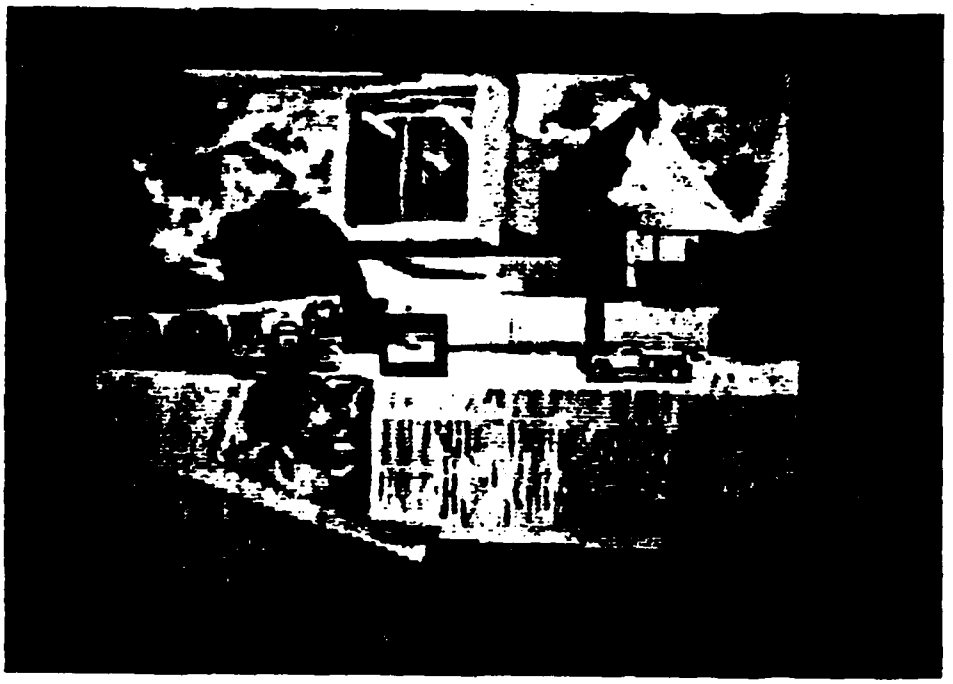
[15] E.C. Hildreth and N.M. Grzywacz, "The incremental recovery of structure from motion: Position vs. velocity based formulations," *Proceedings Workshop on Motion: Representation and Analysis*, May 1986.

[16] J.K. Kearney and W.B. Thompson, "Gradient based estimation of disparity," *Proc. IEEE Conf. on Pattern Recognition and Image Processing*, June, 1982.

[17] "Inexact vision," *Proceedings Workshop on Motion: Representation and Analysis*, May 1986.



(a) Frame 1.



(b) Frame 2.

Figure 3: Image sequence of an indoor scene.

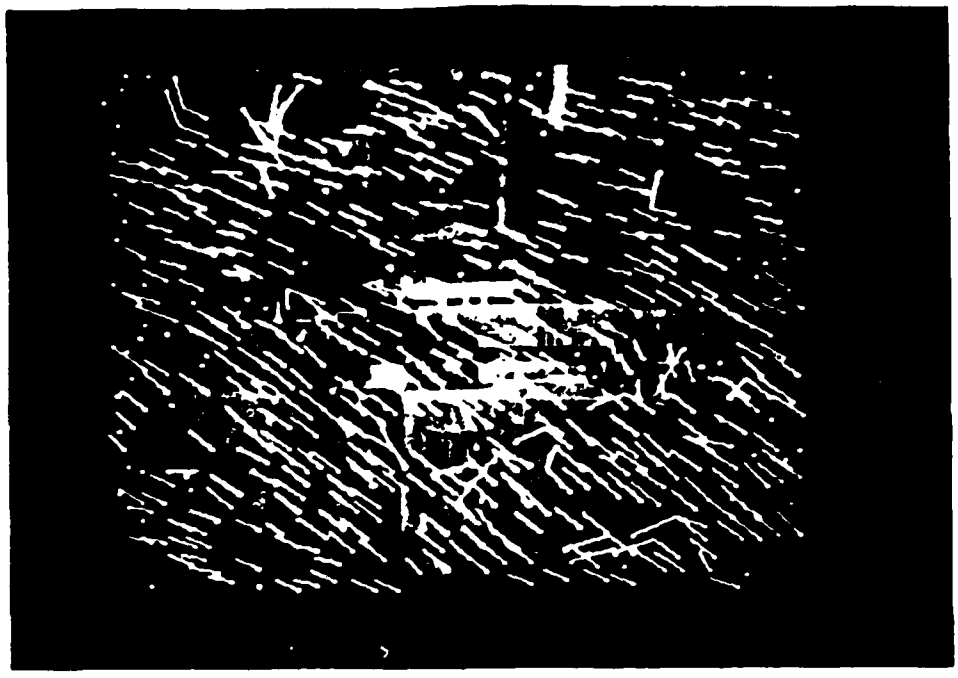


Figure 4: Optical flow field obtained from the image sequence of figure 3.

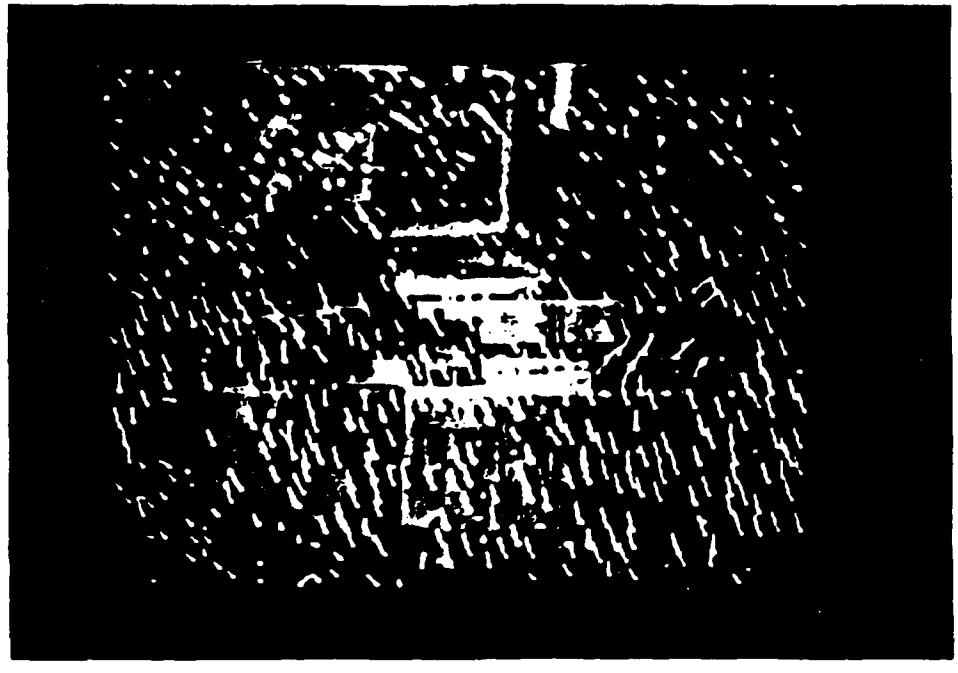


Figure 5: Translational flow field determined from the optical flow field of figure 4.

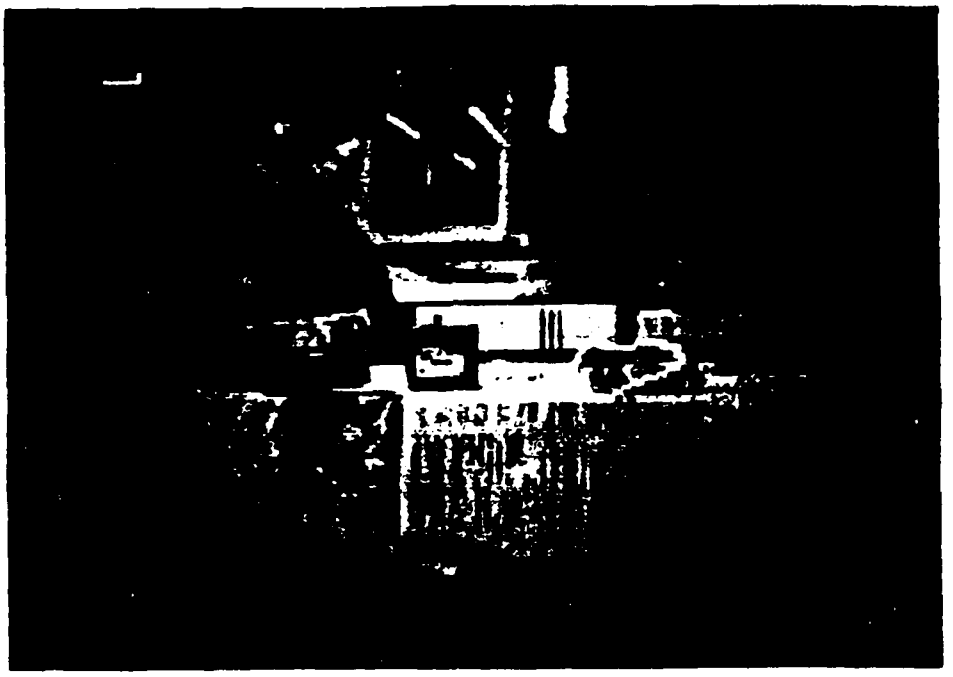
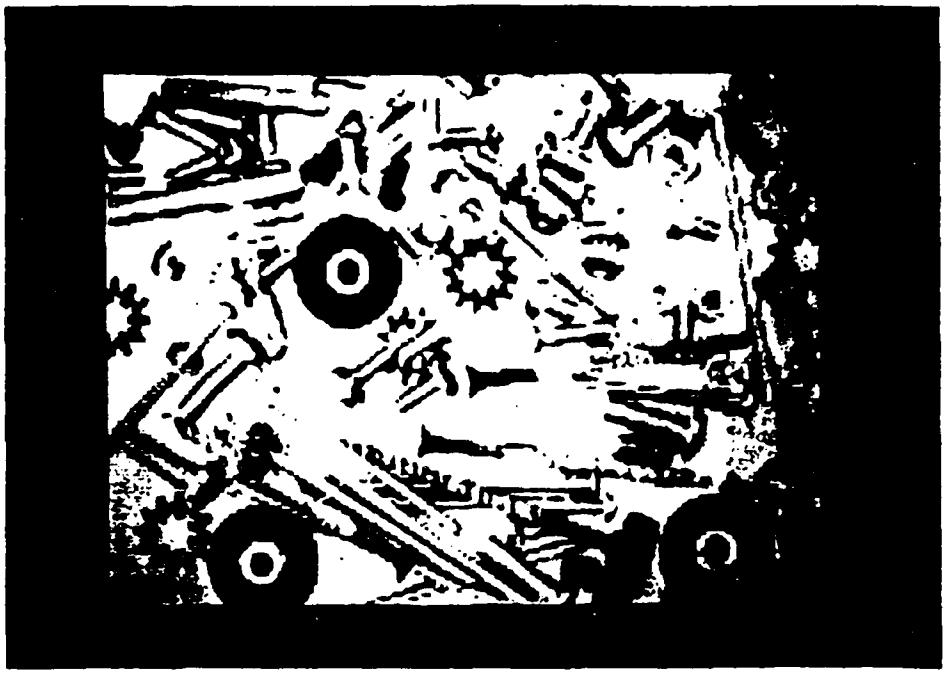
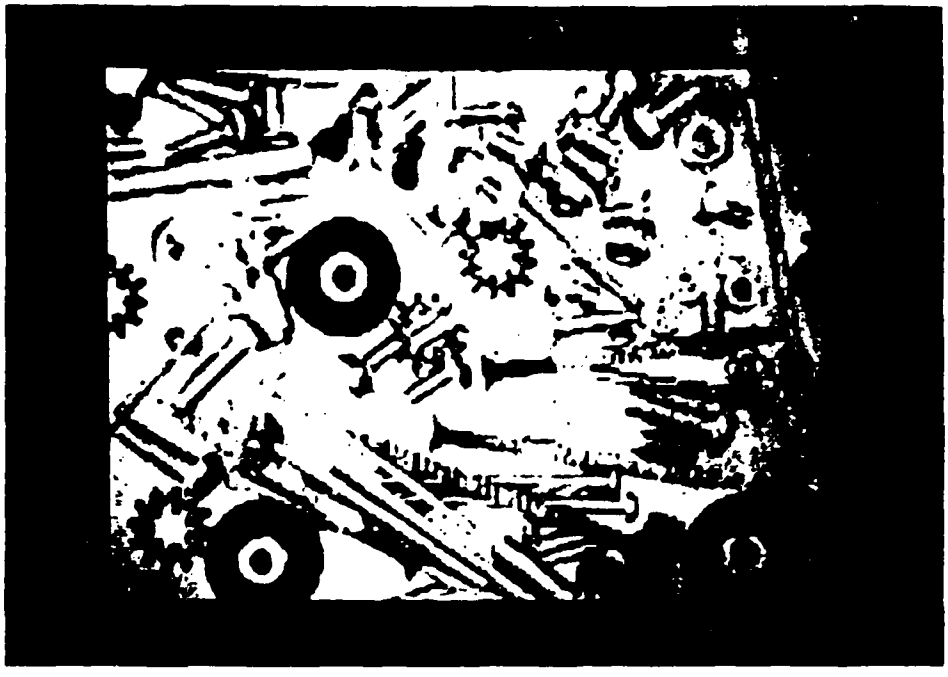


Figure 6: Boundary of a moving object overlaid onto the first image of figure 3.



(a) Frame 1.



(b) Frame 2.

Figure 7: Image sequence of nuts and bolts images.

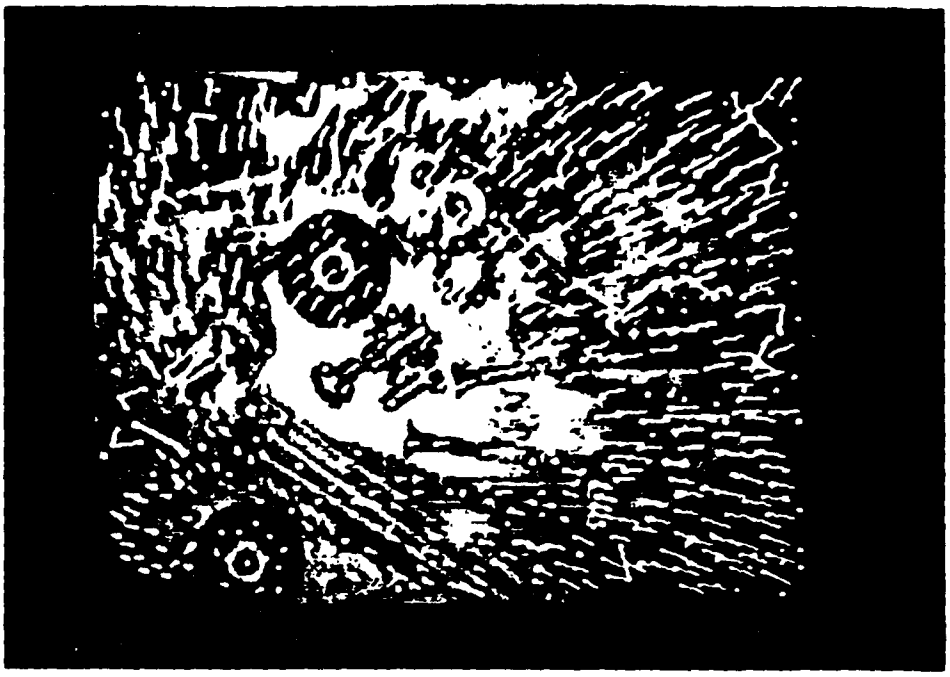
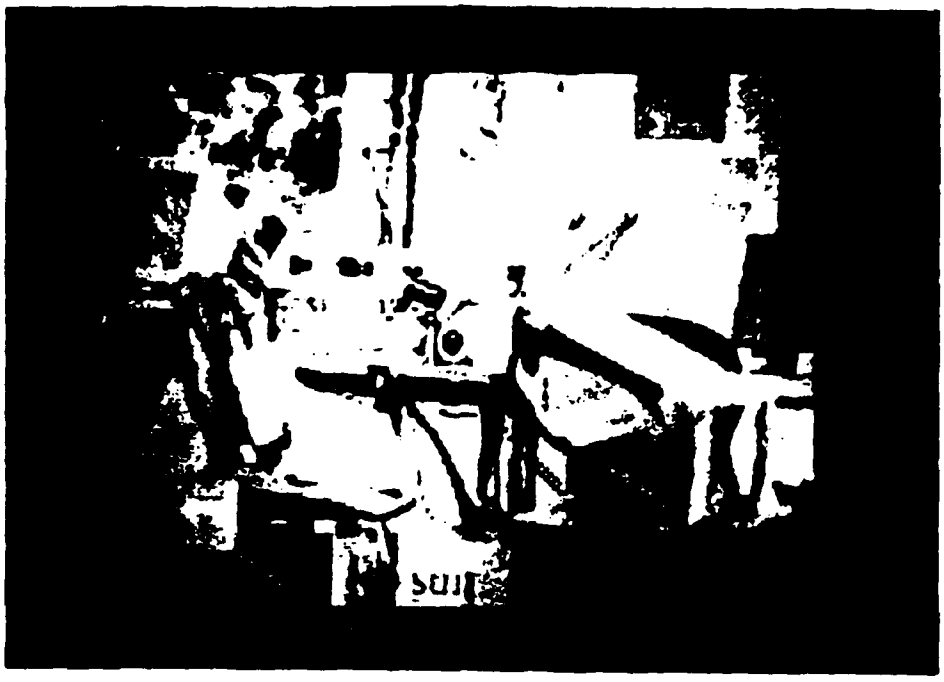
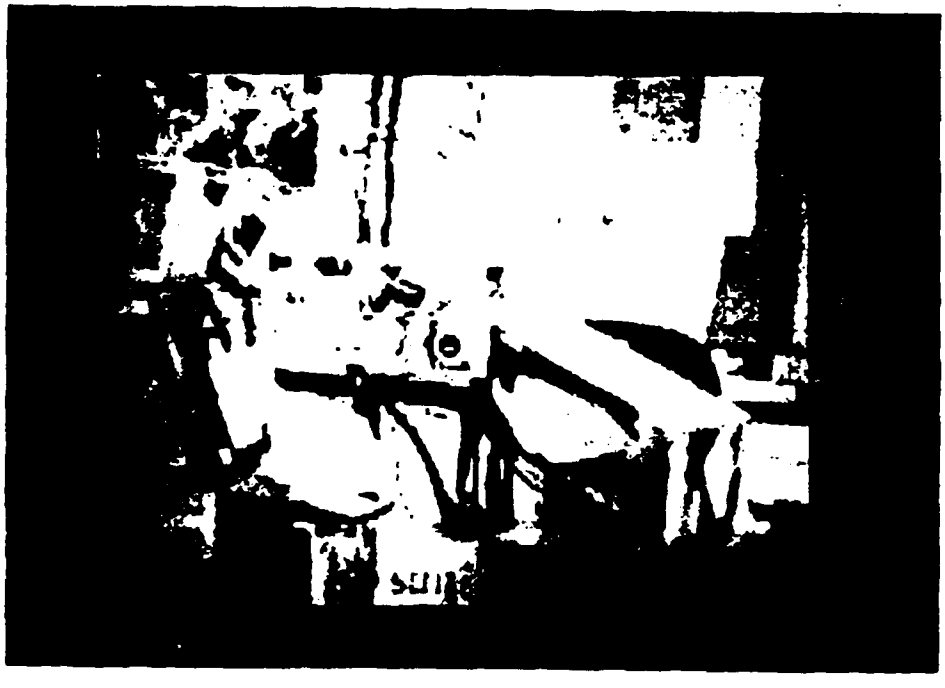


Figure 8: Optical flow field obtained from the image sequence of figure 7.

Figure 9: Boundary of a moving object overlaid onto the first image of figure 7.



(a) Frame 1.



(b) Frame 2.

Figure 10: Image sequence of an indoor scene.

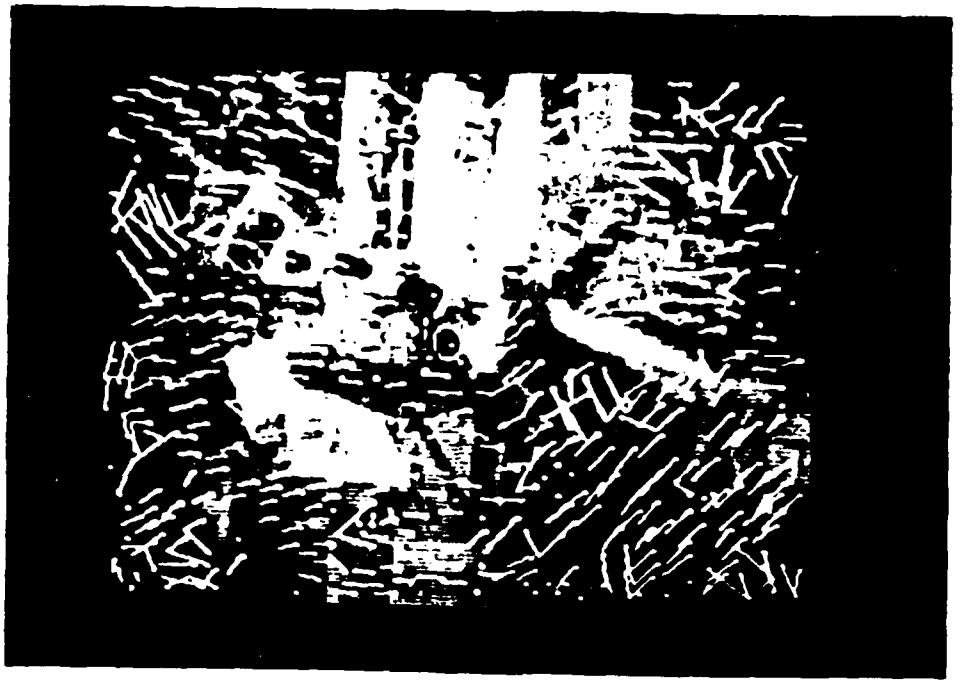


Figure 11: Optical flow field obtained from the image sequence of figure 10.

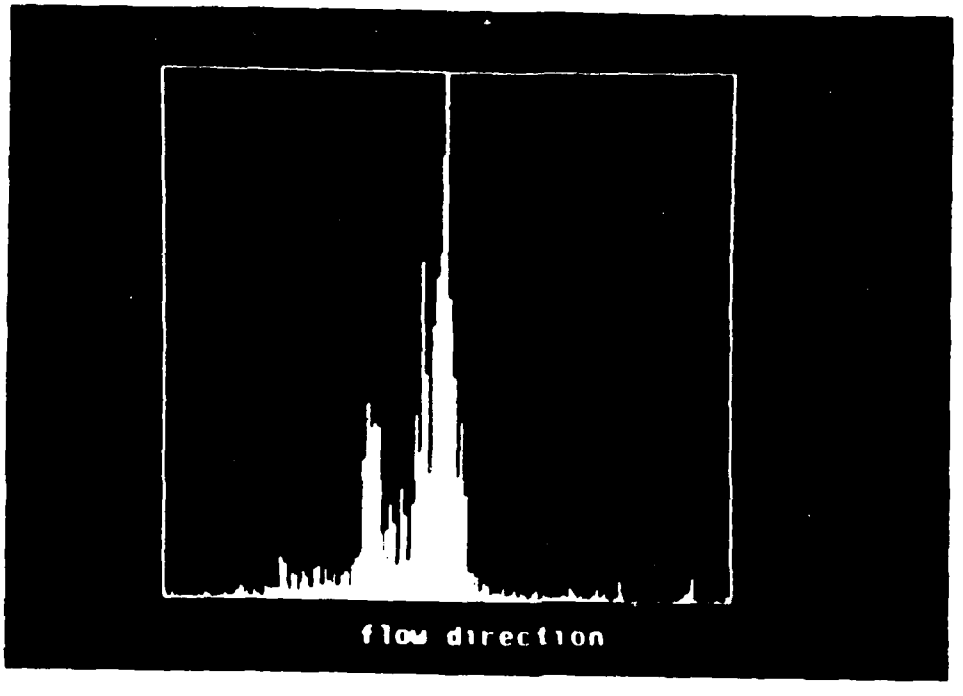


Figure 12: Histogram of the flow directions of the optical flow vectors in figure 10.

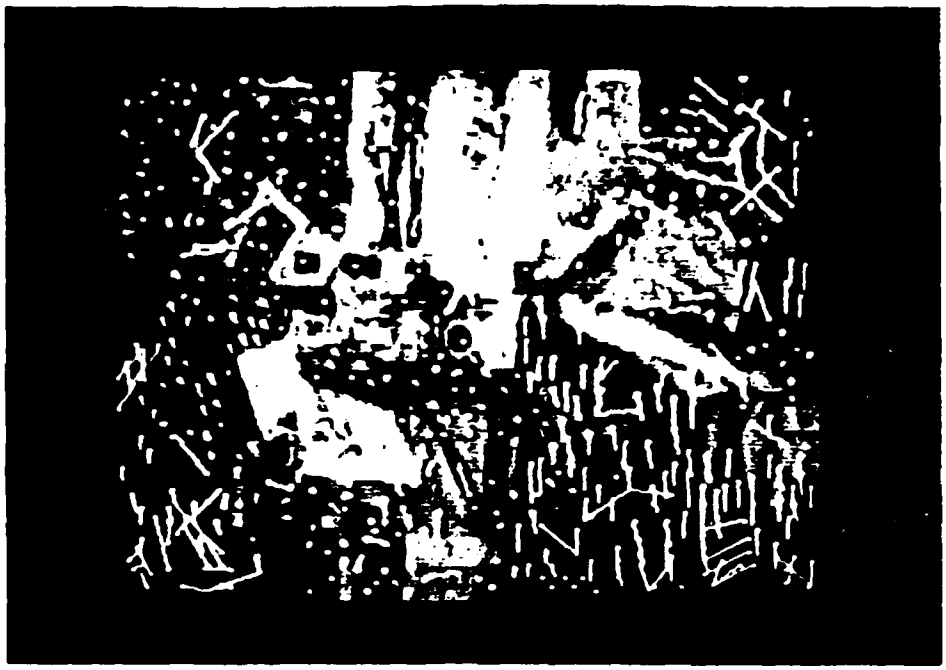


Figure 13: Optical flow field obtained from tracking an object which is stationary with respect to the environment.

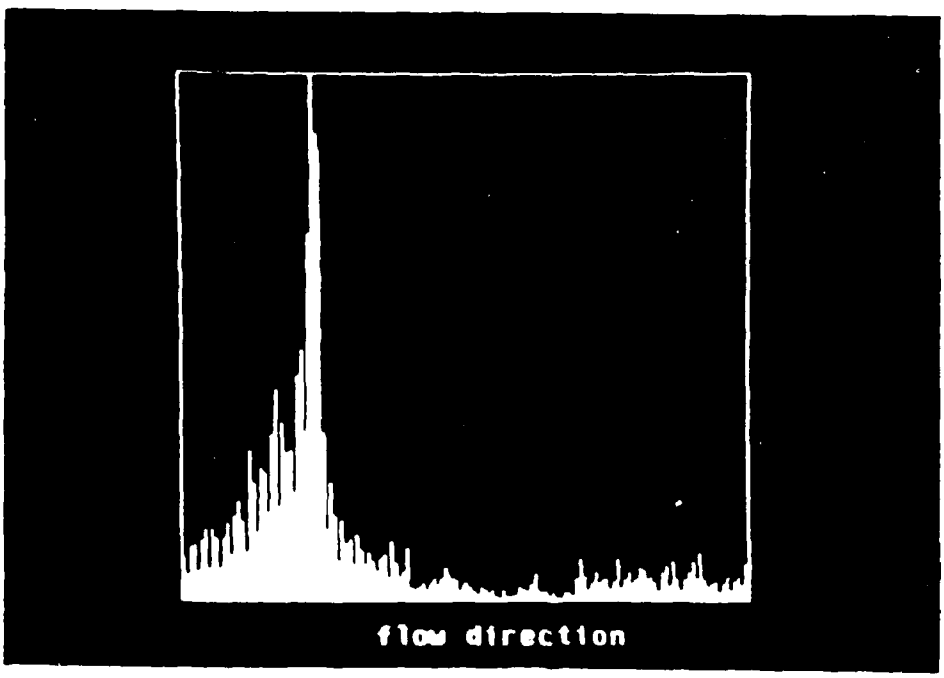


Figure 14: Histogram of the flow directions of the optical flow vectors in figure 13.

Appendix IV -- Acceleration-based structure from motion (Introduction).

The structure from motion problem is inherently ill-conditioned. As a result, present methods for solving the problem are unstable in the presence of noise.

Solution methods may be considered to have three parts: the *information* which they require to be measured from the scene; the *constraints* which are placed on the possible interpretation of the data; and the actual *algorithm* which derives an interpretation from the data and constraints.

Current research has focused on improving algorithms or on adding constraints in order to reduce sensitivity to noise. Much work has actually been devoted to *decreasing* the amount of information gathered from the scene. When more information is used, it is normally only an increase in the number of points examined in the scene. Unfortunately, this requires adding the constraint that the new points are on the same object as the old points. This is a form of spatial continuity assumption and is difficult to enforce. It requires either a perfect segmentation of the scene or a search through the set of possible groupings of points to objects, the combinatorics of which are problematic.

We wish to find new *types* of information to exploit. In particular, we wish to find types of information which can be exploited without adding new constraint to the world or at least by adding very little constraint.

In this paper we will explore the use of derivatives of motion, particularly acceleration, in solving the structure from motion problem. We will show that the problem can be solved from a single point given sufficient (3) derivatives of motion. In practice, we will not be able to accurately estimate high-order derivatives of motion so we will also develop a method which uses only velocity and acceleration but which integrates information from many points. Acceleration constrains the possible interpretations of a point to a one-dimensional family, thus allowing the use of a clustering algorithm. Clustering algorithms do not require a priori knowledge of the grouping of image points to objects, thus removing the difficulty of integrating information from multiple points. Finally, acceleration removes the ambiguity between translation and rotation of an object, thus allowing the object motion to be expressed in the most natural coordinate system.

END

12-87

DTIC